Streamlining cardiopulmonary exercise testing for use as a screening and tracking tool in primary care

Chul-Ho Kim[1], Erik H. Van Iterson[1], James E. Hansen[2], Dean J. MacCarter[1] and Bruce D. Johnson[1]

[1]Department of Cardiovascular Diseases, Mayo Clinic, Rochester, MN, USA; [2]Respiratory and critical care physiology and medicine, Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Torrance, CA, USA

Abstract

Cardiopulmonary exercise testing (CPET) using a spectrum of different approaches demonstrates usefulness for objectively assessing patient disease severity in clinical and research settings. Still, an absence of trained specialists and/or improper data interpretation techniques can pose major limitations to the effective use of CPET for the clinical classification of patients. This study aimed to test an automated disease likelihood scoring algorithm system based on cardiopulmonary responses during a simplified step-test protocol. For patients with heart failure (HF), pulmonary hypertension (PAH), obstructive lung disease (OLD), or restrictive lung disease (RLD), we compared patient scores stratified into one of four "silos" generated from our novel algorithm system against patient evaluations provided by expert clinicians. Patients with HF (n = 12), PAH (n = 9), OLD (n = 16), or RLD (n = 10) performed baseline pulmonary function testing followed by submaximal step-testing. Breath-by-breath measures of ventilation and gas exchange, in addition to oxygen saturation and heart rate were collected continuously throughout testing. The algorithm demonstrated close alignment with patient assessments provided by clinical specialists: HF (r = 0.89, P < 0.01); PAH (r = 0.88, P < 0.01); OLD (r = 0.70, P < 0.01); and RLD (r = 0.88, P < 0.01). Furthermore, the algorithm was capable of differentiating major disease from other disease pathologies. Thus, in a clinically relevant manner, these data suggest this simplified automated disease algorithm scoring system used during step-testing to identify the likelihood that patients have HF, PAH, OLD, or RLD closely correlates with patient assessments conducted by trained clinicians.

Introduction

Comprehensive cardiopulmonary exercise testing (CPET) has traditionally been offered as a clinical diagnostic tool in large-to-mid-sized medical centers. It is typically conducted as a test aimed at pushing patients to maximal exertion, which includes acquisition of physiological data via 12-lead ECG, pulse oximetry, blood pressure, and metabolic cart systems. For successful completion of CPET this commonly requires two trained specialists with oversight by a physician, which can often take 1 h or longer for preparation, testing, cool down, and discharge. In addition, the majority of commercially available metabolic cart systems display an array and sometimes daunting number of cardiopulmonary measurements, often displayed breath-by-breath, which typically requires a higher level of expertise for test interpretation.[1] Thus, with the overall complexity of CPET requiring specialized personnel and time involvement, modest-to-moderate patient risk, and the need for technological integration of multiple tools, the cumulative effect of these factors can act as a barrier for the routine implementation of this powerful clinical assessment resource.

Corresponding author:
Chul-Ho Kim, Mayo Clinic, Department of Cardiovascular Diseases, 200 1st Street, SW Rochester, MN 55905-0001, USA.
Email: kim.chulho@mayo.edu

Previous reports from our group have demonstrated that various simplified forms of submaximal exercise testing can be used to improve the pathophysiological understanding of abnormal breathing patterns and gas exchange in patients with heart failure (HF) and/or pulmonary hypertension (PAH).[2–6] As an important next step for this line of research,[7] we propose there is immediate clinical utility in developing a simplified disease likelihood algorithm based on data acquired via a combination of rest and submaximal exercise physiological testing. Output from our novel algorithm would consider patient responses from a basic resting forced vital capacity (FVC) maneuver in addition to continuous breath-by-breath measurements from a sequence involving 2 min of rest, a 3-min progressive step test, and 1 min of recovery. The overarching goal of such a system is to simplify clinical exercise testing while alleviating burdens such as complex data interpretation.

To determine how well our disease likelihood algorithm performs in properly identifying different cardiopulmonary disease types, we recruited adult patients demonstrating a primary diagnosis of restrictive lung disease (RLD), obstructive lung disease (OLD), chronic HF, or PAH. Patient outputs from algorithm scoring provided ranks for the likelihood of a given disease, which were then compared with separate patient scoring provided by expert clinician reviewers (i.e. professionals whose primary responsibility is the interpretation of clinical exercise tests). Accordingly, this study tested the hypotheses that: (1) scores provided by our novel disease likelihood scoring algorithm would align with those provided by clinicians; and (2) algorithm scores can be used to properly categorize patients in hierarchical order with respect to the most likely primary diagnosis.

Methods

Study design and patients

To test our study aims, 12 patients with HF, nine patients with PAH, 16 patients with OLD, and ten patients with RLD were screened and recruited using our medical records system. All aspects of the study protocol were reviewed and approved by the Mayo Clinic institutional review board. Before participating in study testing, all patients voluntarily provided written informed consent. Table 1 demonstrates baseline participant characteristics stratified by disease type.

Patients performed upright seated pulmonary function testing while at rest to assess basic measurements of FVC, forced expiratory volume in 1 s (FEV_1), and the FEV_1/FVC ratio according to ATS guidelines.[8] This was followed by a brief rest period and a submaximal exercise step test. The exercise protocol consisted of three phases including 2 min of standing rest, 3 min of submaximal incremental stepping exercise, and 1 min of recovery. During the exercise phase, step rates were increased every minute (60, 80, and 100 step/min controlled via metronome) equivalent to 15, 20, and 25 steps/min, respectively. During all phases, ventilation and gas exchange were continuously measured breath-by-breath (SHAPE Medical Systems Inc., St. Paul, MN,

Table 1. Participant characteristics and baseline pulmonary function across groups.

Variables	HF (n = 12)	PAH (n = 9)	OLD (n = 16)	RLD (n = 10)	χ^2	P value
Gender (female/male)	8/4	7/2	10/6	1/9		
Age (years)	60 ± 13	50 ± 15	54 ± 12	68 ± 8		
Height (cm)	170.8 ± 9.2	164.4 ± 6.1	169.4 ± 7.8	176.7 ± 7.5		
Weight (kg)	84.8 ± 15.5	82.3 ± 12.5	83.3 ± 14.7	92.9 ± 9.3		
BMI (kg/m^2)	29.3 ± 6.1	30.5 ± 4.9	28.9 ± 3.8	29.9 ± 3.6		
LVEF (%)	42 ± 11	N/A	N/A	N/A		
NYHA (I/II/III/IV)	0/7/4/1	–	–	–		
FVC (L)	2.70 ± 0.94	2.96 ± 0.64	3.26 ± 0.81	2.83 ± 0.77	3.95	0.27
FVC (%pred.)	69 ± 16	80 ± 13	81 ± 12*	64 ± 10	11.42	<0.01
FEV_1 (L)	2.35 ± 0.78	2.54 ± 0.62	2.66 ± 0.82	2.59 ± 0.64	1.21	0.75
FEV_1 (%pred.)	78 ± 17	87 ± 14	85 ± 18	79 ± 11	2.92	0.40
FEV_1/FVC	88 ± 6	85 ± 5	81 ± 11*	92 ± 4	10.29	0.02
FEV_1/FVC (%pred.)	113 ± 6[†]	108 ± 8[‡]	103 ± 12*	124 ± 8	21.13	<0.001

Data presented as means ± SD.

Table P value and χ^2 value (dF = 3) represents Kruskal–Wallis H test for the overall group effect for each variable. Pairwise differences for significant Kruskal–Wallis H tests were assessed using Wilcoxon rank-sum tests.

*$P < 0.05$, OLD vs. RLD.

[†]$P < 0.05$, HF vs. OLD;

[‡]$P < 0.05$, PH vs. RLD.

HF, heart failure; PH, pulmonary hypertension; OLD, obstructive lung disease; RLD, restrictive lung disease; FVC, forced vital capacity; %pred., percent of predicted values; FEV_1, forced expiratory volume in 1 s.

USA). Heart rate (HR) and rhythm via 12-lead ECG as well as oxygen saturation (SpO_2) via forehead pulse oximetry were also continually monitored.

With respect to our study aims, a pathology likelihood algorithm based on breath-by-breath ventilation and gas exchange responses during testing was used to create patient group stratifications or "silos." The silo likelihood that a patient demonstrated a given primary disease was scored on a range of 0–3 commensurate with the likelihood of having HF, PAH, OLD, or RLD, respectively. Presented in Table 2 are key cardiopulmonary variables input into the scoring algorithm in order to derive each unique silo with respect to a given patient group.

In utilizing the silo system, scores representing each of the four disease categories were recorded for each participant. Importantly, the silo algorithm was designed to weigh scores in an incremental manner based on increasing to decreasing likelihood of a participant demonstrating any one of the primary diseases. Additionally, in utilizing the same scoring scale, separate weighted scores for each of the four disease stratifications were recorded for each participant by three separate expert clinicians from the Mayo Clinic Cardiovascular Health Clinic stress testing practice. Thus, clinician scores were considered the "criterion" method, whereas the silo scoring system was considered the "practical" method to be tested for validity and reliability in being able to identify patients with either HF, PAH, OLD, or RLD. Clinician experts in our stress practice typically review over 50 stress tests daily across a wide spectrum of patient disease etiologies.

Statistical analyses

Where applicable, data are presented at means ± SD. Relationships between silo and clinician scores for each group were assessed using Pearson product moment correlation tests. Mean differences between silo and clinician scoring across disease stratifications were tested using two-factor mixed model analysis of variance (ANOVA) models, which included silo-by-clinician interaction terms for each model. In the event of significant interaction terms, Tukey–Kramer post-hoc corrections were performed to identify where between-within significance occurred. Validity was determined using the following indices: standard error of estimate (SEE) with 95% confidence limits (CL); mean bias (mean difference between silo and mean clinician scores) with 95% limits of agreement (LOA, $\pm 1.96 \times SD$ of differences between scores);[9] and Pearson product moment correlation coefficient (r) with 95% CL. We interpreted the magnitude of r values based on thresholds of Cohen[10] as follows: small = 0.10; medium = 0.30; and large = 0.50; whereby larger is better. Interrater reliability was determined by intraclass correlation coefficients ($ICC = \sigma_B^2/(\sigma_B^2 + \sigma_W^2)$, where σ_B^2 is between variance, whereas σ_W^2 is within variance) with 95% CL as well as standard error of measurement ($St.SEM = SD \times \sqrt{(1-ICC)}$) with 95% CL.[11,12] Interpretation of ICC was determined using the following thresholds: poor < 0.40; moderate 0.40 to < 0.74; excellent > 0.75. All two-tailed statistical significance was determined using an alpha level set at 0.05. Statistical analyses were performed using SPSS programming (version 22.0).

Results

All participants were able to complete exercise testing as described above. At end exercise, there were no significant differences for respiratory exchange ratio (HF versus PAH versus OLD versus RLD; 0.92 ± 0.09 versus 0.92 ± 0.06 versus 0.97 ± 0.14 versus 1.05 ± 0.23, respectively; $P > 0.05$), HR (HF versus PAH versus OLD versus RLD; 102 ± 18 versus 115 ± 16 versus 115 ± 14 versus 108 ± 16 bpm, respectively; $P > 0.05$), relative VO_2 (HF versus PAH versus OLD versus RLD; 15.2 ± 4.2 versus 16.6 ± 4.8 versus 16.6 ± 3.7 versus 13.9 ± 4.3 mL/kg/min, respectively; $P > 0.05$), and the ratio of ventilation to maximal voluntary ventilation (HF versus PAH versus OLD versus RLD; 0.43 ± 0.16 versus 0.57 ± 0.15 versus 0.44 ± 0.21 versus 0.51 ± 0.17, respectively; $P > 0.05$).

For baseline pulmonary function, HF and RLD demonstrated significantly lower FVC than PAH and OLD (both $P < 0.01$), whereas there were no significant differences in FEV_1 across groups ($P > 0.05$). Consequently, HF and

Table 2. Key variables in disease pathology likelihood algorithm.

HF silo	PAH silo	OLD silo	RLD silo
VE/VCO_2 slope[14,15]	VE/VCO_2 slope[14,15]	SpO_2 at peak[16]	SpO_2 at peak[16]
O_2 p/VO_2 slope[17]	peak GxCap[13]	FEV_1% pred.[18,19]	FVC% pred.[19]
OUES[20,21]	MPIph[22–24]	$PECO_2$/$PETCO_2$[25]	VTmax/rest[26]
CirEquVO2 [27]	SpO_2 at rest	Breathing res.[16]	Lung stiffness slope[28]
HR rec[29]	SpO_2 at peak[16]		

HF, heart failure; PAH, pulmonary arterial hypertension; OLD, obstructive lung disease; RLD, restrictive lung disease; VE/VCO_2, ventilatory efficiency; O_2p, oxygen pulse: oxygen consumption/heart rate; VO_2, oxygen consumption; OUES, oxygen uptake efficiency slope; CircEq VO_2 % pred, circulatory equivalent oxygen consumption; HR, heart rate; MPIph, multi-parameter index for pulmonary hypertension; GxCap, pulmonary capacitance: oxygen pulse × the partial pressure of end tidal CO_2; SpO_2, oxygen saturation; FEV_1, forced expiratory in 1 s; $PECO_2$, the partial pressure of mean expired CO_2; $PETCO_2$, the partial pressure of end tidal CO_2; FVC % pred, % predicted of forced vital capacity.
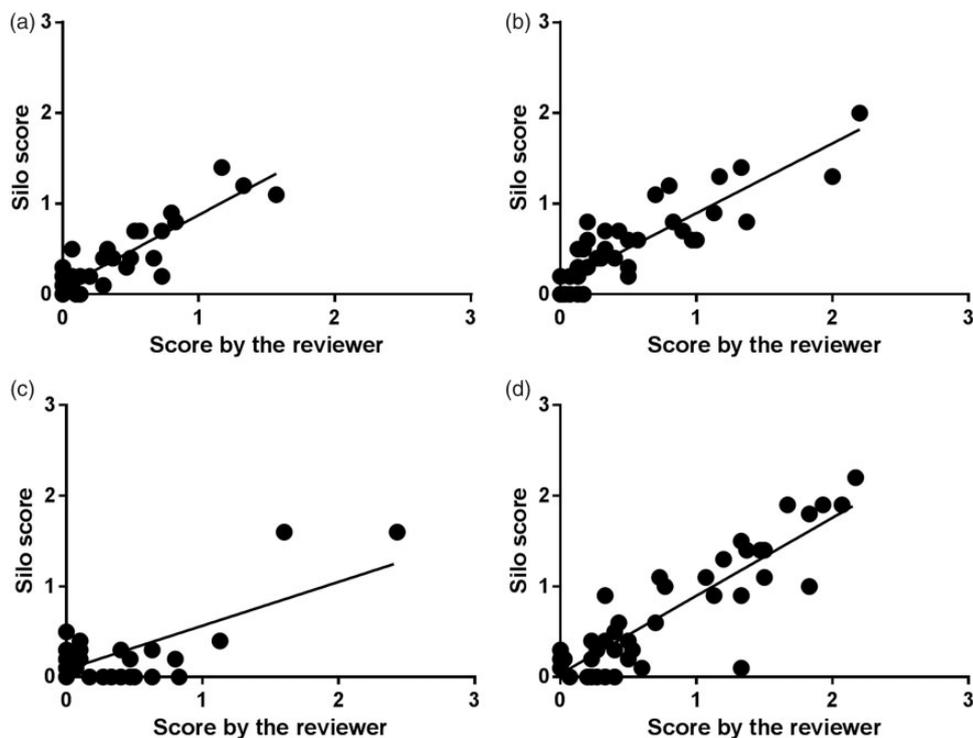
Fig. 1. Relationship between silo score and scores by the reviewers for entire group (n = 47). (a) HF (r = 0.89, P < 0.01); (b) PAH (r = 0.88, P < 0.01); (c) OLD (r = 0.70, P < 0.01); and (d) RLD (r = 0.88, P < 0.01).

RLD demonstrated higher FEV_1/FVC when compared to PAH and OLD ($P < 0.01$, Table 1).

Figure 1 illustrates there was close alignment between silo and clinician scoring for the likelihood of each possible diagnosis. For HF, the score relationship was strong between silo and clinician (r = 0.89, $P < 0.01$; Fig. 1a). Scoring for PAH between silo and clinician was also strong (r = 0.88, $P < 0.01$; Fig. 1b). Likewise, scoring between silo and clinician was strong for both OLD (r = 0.70, $P < 0.01$) and RLD (r = 0.88, $P < 0.01$; Fig. 1c and d, respectively).

Validity outcomes using the silo system (practical method) against mean Clinician scores (criterion method) presented in Table 3 demonstrated an overall large magnitude of validity across silo 1–4 within each stratification. This was accompanied by consistency of validity scoring across indices for the likelihood of predicting scores given by Clinicians for HF, PAH, OLD, and RLD, irrespective of the primary disease stratification.

Likewise, interrater reliability that considered patient scores for each group, which consisted of scores from each Clinician as well as the silo system, demonstrated modest levels of interrater variability both between and within stratification (Table 4). Overall interrater reliability was particularly strong with respect to scores provided for a primary disease associated with a given group (Table 4). This was followed by a general trend where ICC values for at least three out of four secondary disease classes were > 0.65 within each grouping.

Score distribution patterns across diseases between silo and Clinician scores were similar. In addition to being able to properly score the likelihood of a primary disease, each disease silo was capable of discretely differentiating the likelihood of a given secondary disease. As such, the overall score distribution pattern for the likelihood of HF was not different between silo and clinician ($P > 0.05$; Fig. 2a). Similar outcomes were observed for PAH ($P > 0.05$; Fig. 2b), OLD ($P > 0.05$; Fig. 2c), and RLD ($P > 0.05$; Fig. 2d).

Finally, and consistent with observations above, for the HF silo in Fig. 2a, the diagnosis of HF demonstrated the highest score for both silo and Clinicians ($P < 0.05$). Similar results were also observed for PAH ($P < 0.05$; Fig. 2b) and RLD ($P < 0.05$; Fig. 2d). However, OLD did not demonstrate a "highest" score for the OLD silo ($P > 0.05$; Fig. 2c).

Discussion

The present study tested how well a novel automated algorithm (i.e. silo scores), based on variables acquired from a brief and simple cardiopulmonary test, is able to properly differentiate primary and secondary disease types compared with similar assessments made by trained Clinicians. These data suggest that for each silo (i.e. HF, PAH, OLD, and RLD), output scores closely correlated with scores provided by Clinicians, which was accompanied by similar disease score distributions across methods. In general, the present observations suggest our automated cardiorespiratory

Table 3. Validity indices of patient groups scored by clinicians (criterion) vs. silo algorithm (practical).

Group (Silo)	SEE	Std. SEE	r	Bias
HF (n = 12)	0.18 (0.12–0.31)	0.38 (0.27–0.67)	0.93 (0.77–0.98)	−0.06 (−0.18–0.05)
HF (1)	–	–	–	–
PAH (2)	0.21 (0.15–0.36)	0.38 (0.27–0.67)	0.93 (0.77–0.98)	−0.08 (−0.23–0.07)
OLD (3)	0.29 (0.20–0.51)	1.04 (0.73–1.82)	−0.13 (−0.66–0.48)	−0.21 (−0.41−−0.01)
RLD (4)	0.32 (0.22–0.56)	0.44 (0.31–0.78)	0.91 (0.69–0.97)	−0.07 (−0.26–0.13)
PAH (n = 9)	0.36 (0.24–0.74)	0.55 (0.37–1.13)	0.85 (0.44–0.97)	−0.13 (−0.39–0.13)
HF (1)	0.22 (0.15–0.45)	0.45 (0.29–0.91)	0.91 (0.62–0.98)	−0.01 (−0.18–0.16)
PAH (2)	–	–	–	–
OLD (3)	0.44 (0.29–0.89)	1.02 (0.68–2.08)	0.29 (−0.46–0.80)	−0.04 (−0.36–0.27)
RLD (4)	0.23 (0.16–0.48)	0.43 (0.29–0.88)	0.91 (0.63–0.98)	−0.03 (−0.20–0.15)
OLD (n = 16)	0.33 (0.24–0.52)	0.48 (0.35–0.76)	0.88 (0.69–0.96)	−0.11 (−0.29–0.07)
HF (1)	0.15 (0.11–0.23)	0.79 (0.58–1.24)	0.65 (0.23–0.87)	0.06 (−0.02–0.13)
PAH (2)	0.15 (0.11–0.24)	0.50 (0.36–0.79)	0.88 (0.67–0.96)	0.11 (0.00–0.21)
OLD (3)	–	–	–	–
RLD (4)	0.29 (0.21–0.45)	0.83 (0.61–1.31)	0.60 (0.15–0.84)	−0.03 (−0.18–0.13)
RLD (n = 10)	0.37 (0.25–0.72)	0.65 (0.44–1.24)	0.79 (0.32–0.95)	−0.16 (−0.46–0.14)
HF (1)	0.09 (0.06–0.16)	0.75 (0.50–1.43)	0.71 (0.15–0.93)	0.10 (0.00–0.20)
PAH (2)	0.14 (0.09–0.26)	0.32 (0.22–0.61)	0.95 (0.81–0.99)	0.15 (0.06–0.24)
OLD (3)	0.05 (0.03–0.10)	1.05 (0.71–2.02)	0.11 (−0.56–0.69)	0.22 (0.15–0.28)
RLD (4)	–	–	–	–

Validity indices reported with 95% confidence limits in parentheses. Validity indices were calculated using clinician mean scores (criterion) against corresponding silo scores (practical). Rows with dashes correspond with data in the top for row for each patient group indicating the respective silo (no.) that was set to predict the primary likelihood of that patient condition (e.g. for HF, silo 1 indicated in parentheses as (1), was set to primarily predict the likelihood of HF, and therefore validity data for that Group-Silo pairing has been placed on the top row for patients with HF). There are validity scores for each condition within a given patient stratification because irrespective of patient group, both clinicians and silo algorithm scored patients for the likelihood of having each of the four conditions. HF, heart failure; PAH, pulmonary hypertension; OLD, obstructive lung disease; RLD, restrictive lung disease; SEE, standard error of estimate; Std. SEE, standardized standard error of estimate; r, Pearson product moment correlation coefficient; bias, mean bias = mean difference between values.

Table 4. Interrater reliability indices of patient groups scored by clinicians and silo algorithm.

Group (Silo)	ICC	St. EM	Std. St. EM
HF	0.81 (0.57–0.94)*	0.25 (0.20–0.35)	0.50 (0.39–0.70)
HF (1)	–	–	–
PAH (2)	0.93 (0.82–0.98)*	0.18 (0.14–0.25)	0.31 (0.24–0.44)
OLD (3)	0.42 (0.07–0.82)†	0.31 (0.24–0.45)	0.81 (0.63–1.19)
RLD (4)	0.86 (0.69–0.95)*	0.31 (0.25–0.44)	0.41 (0.33–0.57)
PAH	0.90 (0.74–0.97)*	0.25 (0.19–0.37)	0.37 (0.28–0.55)
HF (1)	0.98 (0.92–0.99)*	0.13 (0.10–0.20)	0.25 (0.19–0.38)
PAH (2)	–	–	–
OLD (3)	0.47 (0.09–0.82)†	0.40 (0.31–0.61)	0.76 (0.59–1.15)
RLD (4)	0.98 (0.93–0.99)*	0.11 (0.09–0.17)	0.21 (0.16–0.31)
OLD	0.88 (0.72–0.97)*	0.30 (0.24–0.41)	0.41 (0.32–0.57)
HF (1)	0.47 (0.05–0.91)†	0.23 (0.17–0.38)	0.82 (0.60–1.34)
PAH (2)	0.74 (0.50–0.92)*	0.19 (0.15–0.25)	0.55 (0.44–0.75)
OLD (3)	–	–	–
RLD (4)	0.83 (0.64–0.94)*	0.17 (0.14–0.23)	0.46 (0.37–0.61)

(continued)

Table 4. Continued

Group (Silo)	ICC	St. EM	Std. St. EM
RLD	0.96 (0.89–0.99)*	0.14 (0.11–0.20)	0.24 (0.18–0.34)
HF (1)	0.68 (0.36–0.90)*	0.08 (0.06–0.12)	0.61 (0.47–0.87)
PAH (2)	0.84 (0.64–0.95)*	0.20 (0.15–0.29)	0.44 (0.34–0.63)
OLD (3)	0.35 (0.00–0.73)†	0.06 (0.05–0.09)	0.83 (0.64–1.19)
RLD (4)	–	–	–

Reliability indices reported with 95% confidence limits in parentheses. Interrater reliability indices were calculated using each clinician score as well as corresponding silo score. Rows with dashes correspond with data in the top for row for each patient group indicating the respective silo (no.) that was set to predict the primary likelihood of that patient condition (e.g. for HF, silo 1 indicated in parentheses as (1), was set to primarily predict the likelihood of HF, and therefore reliability data for that Group-Silo pairing has been placed on the top row for patients with HF). There are reliability scores for each condition within a given patient stratification because irrespective of patient group, both clinicians and silo algorithm scored patients for the likelihood of having each of the four conditions.
*P < 0.0001.
†P < 0.05.
HF, heart failure; PAH, pulmonary hypertension; OLD, obstructive lung disease; RLD, restrictive lung disease; ICC, intraclass correlation coefficient; St. EM, standard error of measurement; Std. St. EM, standardized standard error of measurement.
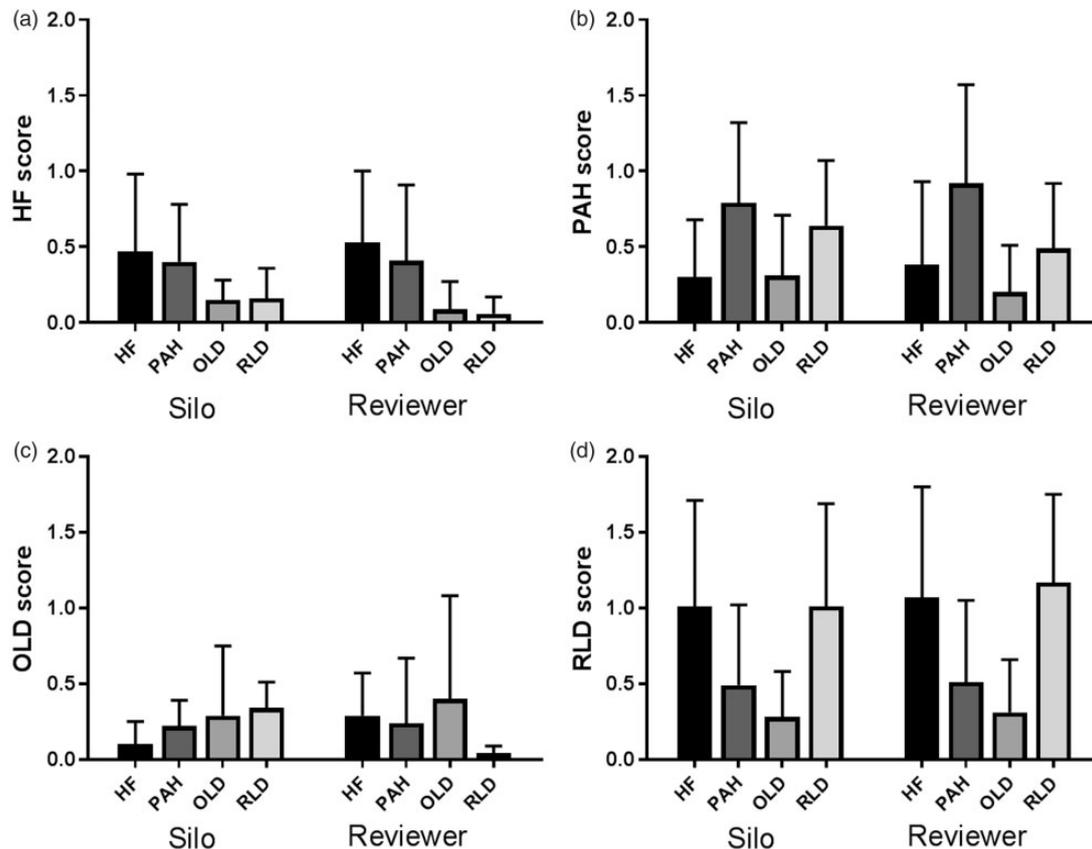
Fig. 2. The pattern of score distribution across disease in silo and the reviewers: HF; PAH; OLD; and RLD.

algorithm consistently demonstrated close scoring alignment with primary patient disease class stratifications identified by experienced Clinicians.

The observations from this study support the proposed intent of our original disease likelihood/severity scoring algorithm, which was designed to help with simplifying exercise testing implementation and interpretation for routine use in clinical and/or laboratory settings. These data are consistent with our previous reports showing that for a spectrum of cardiopulmonary diseases, clinical assessment of functional capacity does not always require maximal exercise testing.[2,5,6,13] The present observations suggest that changes in key cardiopulmonary measurements can be detected with relatively mild to moderate levels of physical exertion. An additional strength of our proposed algorithm system is that when conducting submaximal exercise testing, the risk of patient events may be reduced leading to the need for less strategic oversight and required in-room personnel. Lastly, the physiological assessment tool in the present study illustrates the ease whereby clinical exercise testing may be able to be routinely performed in the absence of needing to perform time consuming data post-processing and advanced computations. Thus, with the present system, it may be possible to reduce the "extra layer" of interpretive expertise needed.

Our findings suggest that an automated scoring algorithm based on patterns of cardiopulmonary responses from a modified step-test exercise protocol of mild to moderate physical exertion may be used to track common patient clinical phenotypes including HF, PAH, RLD, or OLD. Based on our sample size, we acknowledge the scope of our observations is preliminary in being able to determine the wide-scale real-world clinical efficacy of our novel approach. Still, these data suggest there is immediate potential associated with the present simplified testing paradigm applied to less traditional clinical settings where providing rapid feedback can help with guiding clinical decision making and a simpler way to track patients with chronic diseases to determine need for more expert intervention.

Although it is well-recognized that diseases of the cardiopulmonary system are not exclusively centralized to cardiac versus pulmonary limitations in patients, it remains a challenge for clinicians to routinely and quickly differentiate contributions from primary versus secondary underlying disease processes, such as is the case for example, in patients with HF, PAH secondary to HF, or PAH without HF. Therefore, it is promising that while frequently sharing similar signs and symptoms, these data suggest the current approach may be helpful for separating out conditions that align closely. Moreover, it is important to highlight that the proposed value of this clinical exercise testing method is using a pseudo "activities of daily living" exercise setting (e.g. climbing stairs) where signs and symptoms of

the present diseases are well-acknowledged to be commonly exacerbated in patients.

In contrast to the proposed value of the present exercise testing method aimed at identifying primary disease in patients with HF, PAH, OLD, or RLD, we acknowledge that calibration of sub-condition OLD scoring with clinician scores did not consistently align within each strata. This was followed by an inability of the silo algorithm to consistently and clearly identify OLD from other diseases. Therefore, our future direction will be to study a larger patient sample, which will include a broader spectrum of OLD severity that will prove to strengthen the capability of our algorithm to identify this specific disease. Nevertheless, despite this study limitation, it is important to note that the chronic conditions tested in this study do not tend to exist by themselves and, hence, a clear distinction in patients with variable mild or modest disease is more difficult. With this, we suggest this may be a strength of our algorithm since output scores provide a more inclusive picture of "reality" rather than trying to suggest there is a single limiting factor during activity.

Clinical implications

Clinical CPET produces a complex array of measurements that can be integrated into algorithms allowing for a more simplified approach to screen and track patients. Ultimately, with simplified protocols, this allows CPET to move closer to a "point of care" approach or essentially a vital sign that could be pursued in primary care settings guiding the need and direction for sending patients to subspecialty clinics.

Conflict of interest

The author(s) declare that there is no conflict of interest.

Funding

In memoriam

We are grateful to Dr. James 'Jim' E. Hansen, Professor Emeritus at UCLA Harbor for his contributions on this study. Dr. Hansen was supportive of sub-maximal exercise gas exchange testing to advance its clinical application. James Hansen passed away in May of 2017. He has authored numerous publications and a textbook series on cardiopulmonary testing and its clinical utility in heart, lung and pulmonary vascular disease.

References

1. Johnson B, Whipp B, Zeballos R, et al. Conceptual and physiological basis of cardiopulmonary exercise testing measurement. In the ATS/ACCP statement on cardiopulmonary exercise testing. *Am J Resp Crit Care* 2003; 167: 211–277.
2. Arena R, MacCarter D, Olson TP, et al. Ventilatory expired gas at constant-rate low-intensity exercise predicts adverse events and is related to neurohormonal markers in patients with heart failure. *J Card Fail* 2009; 15: 482–488.
3. Kim C-H, Anderson S, Maccarter D, et al. A multivariable index for grading exercise gas exchange severity in patients with pulmonary arterial hypertension and heart failure. *Pulm Med* 2012; 2012: 962598.
4. Kim C-H, Cha Y-M, Shen W-K, et al. Effects of atrioventricular and interventricular delays on gas exchange during exercise in patients with heart failure. *J Heart Lung Transplant* 2014; 33: 397–403.
5. Woods PR, Bailey KR, Wood CM, et al. Submaximal exercise gas exchange is an important prognostic tool to predict adverse outcomes in heart failure. *Eur J Heart Fail* 2011; 13: 303–310.
6. Woods PR, Frantz RP, Taylor BJ, et al. The usefulness of submaximal exercise gas exchange to define pulmonary arterial hypertension. *J Heart Lung Transplant* 2011; 30: 1133–1142.
7. Kim C-H, Hansen JE, MacCarter DJ, et al. Algorithm for predicting disease likelihood from a submaximal exercise test. *Clin Med Insights Cric Respir Pulm Med* 2017; 11: 1179548417719248.
8. Miller MR, Hankinson J, Brusasco V, et al. Standardisation of spirometry. *Eur Respir J* 2005; 26: 319–338.
9. Hopkins WG, Marshall SW, Batterham AM, et al. Progressive statistics for studies in sports medicine and exercise science. *Med Sci Sports Exerc* 2009; 41: 3–13.
10. Cohen J. A power primer. *Psychol Bull* 1992; 112: 155–159.
11. McGraw KO and Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Meth* 1996; 1: 30.
12. Eliasziw M, Young SL, Woodbury MG, et al. Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. *Phys Ther* 1994; 74: 777–788.
13. Taylor BJ, Olson TP, Chul Ho K, et al. Use of noninvasive gas exchange to track pulmonary vascular responses to exercise in heart failure. *Clin Med Insights Circ Respir Pulm Med* 2013; 7: 53–60.
14. Arena R, Myers J, Abella J, et al. Development of a ventilatory classification system in patients with heart failure. *Circulation* 2007; 115: 2410–2417.
15. Kim C-H, Olson LJ, Shen WK, et al. Ventilatory gas exchange and early response to cardiac resynchronization therapy. *J Heart Lung Transplant* 2015; 34: 1430–1435.
16. Forman DE, Myers J, Lavie CJ, et al. Cardiopulmonary exercise testing: relevant but underused. *Postgrad Med* 2010; 122: 68–86.
17. Wasserman K, Hansen J, Sue D, et al. *Principles of Exercise Testing and Interpretation: Including Pathophysiology and Clinical Applications*. Philadelphia. PA: Lippincott, Williams & Wilkins, 2005.
18. Lindberg A, Larsson LG, Ronmark E, et al. Decline in FEV1 in relation to incident chronic obstructive pulmonary disease in a cohort with respiratory symptoms. *COPD* 2007; 4: 5–13.
19. Hankinson JL, Odencrantz JR and Fedan KB. Spirometric reference values from a sample of the general US population. *Am J Resp Crit Care* 1999; 159: 179–187.
20. Davies LC, Wensel R, Georgiadou P, et al. Enhanced prognostic value from cardiopulmonary exercise testing in chronic heart failure by non-linear analysis: oxygen uptake efficiency slope. *Eur Heart J* 2006; 27: 684–690.
21. Hollenberg M and Tager IB. Oxygen uptake efficiency slope: an index of exercise performance and cardiopulmonary reserve requiring only submaximal exercise. *J Am Coll Cardiol* 2000; 36: 194–201.

22. Matsumoto A, Itoh H, Eto Y, et al. End-tidal CO2 pressure decreases during exercise in cardiac patients: association with severity of heart failure and cardiac output reserve. *J Am Coll Cardiol* 2000; 36: 242–249.

23. Ramos RP, Alencar MCN, Treptow E, et al. Clinical usefulness of response profiles to rapidly incremental cardiopulmonary exercise testing. *Pulmonary Medicine* 2013; 2013: 359021.

24. Sun X-G, Hansen JE, Oudiz RJ, et al. Gas exchange detection of exercise-induced right-to-left shunt in patients with primary pulmonary hypertension. *Circulation* 2002; 105: 54–60.

25. Hansen JE, Ulubay G, Chow BF, et al. Mixed-expired and end-tidal CO2 distinguish between ventilation and perfusion defects during exercise testing in patients with lung and heart diseases. *Chest* 2007; 132: 977–983.

26. Olson TP and Johnson BD. Influence of cardiomegaly on disordered breathing during exercise in chronic heart failure. *Eur J Heart Fail* 2011; 13: 311–318.

27. Sun X-G, Hansen JE and Stringer WW. Oxygen uptake efficiency plateau best predicts early death in heart failure. *Chest* 2012; 141: 1284–1294.

28. Agostoni P, Pellegrino R, Conca C, et al. Exercise hyperpnea in chronic heart failure: relationships to lung stiffness and expiratory flow limitation. *J Appl Physiol* 2002; 92: 1409–1416.

29. Kubrychtova V, Olson TP, Bailey KR, et al. Heart rate recovery and prognosis in heart failure patients. *Eur J Appl Physiol* 2009; 105: 37–45.